DOCUMENT RESUME

ED 174 674 **TM 009 646**

AUTHOR	Schaeffer, Gary A.; And Others
TITLE	A Multitrait-Multimethod Approach to Evaluating Measurement Techniques in Developmental Research.
PUB DATE	Apr 79
NOTE	14p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Behavior Rating Scales; *Child Development; Correlation; Developmental Psychology; Individual Tests; *Matrices; Measurement Goals; *Observation; Preschool Education; Psychological Tests; Questionnaires; *Test Reliability; *Test Selection; Test Validity
IDENTIFIERS	Construct Validity; *Multitrait Multimethod Techniques

ABSTRACT

Developmental psychologists have not traditionally documented the psychometric properties of their observational instruments, partly because such concepts as reliability and validity are associated mainly with paper-and-pencil tests. A multitrait-multimethod correlation matrix (MTMM) was presented as a useful method for the psychometric evaluation of these instruments; two hypothesized constructs important in early social development (fear of strangers and exploration in a novel environment) were measured by a laboratory observation of child behavior, and by a maternal questionnaire. The MTMM applied an evaluation sequence not usually followed by researchers--the sequence consisted of intra- and inter-observer agreement; stability--consistency of children's behavior over a short term; and construct validity (and its prerequisites, convergent and discriminant validity). Observer agreement was a necessary but not sufficient condition for stability, and stability was a necessary but not sufficient condition for validity. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174674

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Multitrait-Multimethod Approach to Evaluating Measurement Techniques in Developmental Research*

Gary A. Schaeffer, Lynne E. Gray, and Mark Everson
School of Education
Stanford University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gary A. Schaeffer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

This paper deals with the necessity of including procedures for the psychometric assessment of observational instruments in child development research. This assessment should include observer agreement, stability, and construct validity. Evaluating the quality of measures is crucial in the accumulation of a sound knowledge base that supports theoretical constructs. The multitrait-multimethod framework is presented as a particularly useful tool for guiding the psychometric evaluation of observational instruments.

IM009 646

* This paper was prepared for the meeting of the American Educational Research Association, San Francisco, April, 1979. Correspondence should be sent to: Gary Schaeffer, Dept. of Psychological Studies, School of Education, Stanford University, Stanford, CA 94305.

Introduction

This paper will suggest the multitrait-multimethod (MTMM) framework for evaluating the quality of observational instruments used in developmental research. Child development researchers are primarily concerned with substantive issues. This emphasis has too often overshadowed concern for the quality of measurement procedures. This is particularly problematic because of the increasing reliance on both controlled and naturalistic observation techniques whose reliability and validity have not been established. Observational instruments must be shown to produce consistent information before they can be relied upon as indicators of developmental constructs. Without psychometrically sound procedures, a dependable knowledge base can not be built up and theories cannot be supported.

There are two main reasons why developmental psychologists have not traditionally tested and documented the psychometric qualities of their observational instruments. First, concepts such as reliability, and especially validity, are usually thought of as applying to paper-and-pencil tests rather that observational techniques. Although traditional evaluative procedures often are not appropriate for observational instruments, the principles and intentions underlying reliability and validity certainly apply and methods to assess them can be implemented. Second, psychometric assessment is often thought to be a particularly complex and costly undertaking. This notion, however, is not necessarily true, especially if at least some assessment can be done in each study and data from different studies accumulate.

Many researchers indeed do investigate the psychometric properties of

their instruments, but frequently the reporting and documentation of results is unsystematic or inadequate. There is a logical sequence of techniques for evaluating observational instruments that is usually not followed by researchers. The sequence consists of observer agreement, stability, and validity, where observer agreement is a necessary but not sufficient condition for stability, and stability is a necessary but not sufficient condition for validity (Frick and Semmel, 1978). The Campbell and Fiske (1959) MTMM orientation can guide the collection of psychometric data and structure the researcher's evaluation of assessment techniques so that the advancement of knowledge in developmental research can be more systematic, coherent, and cumulative.

The following simplified example of a hypothetical MTMM correlation matrix will serve as an introduction.

	FEAR OF STRANGERS		EXPLORATION	
	1. Lab Observation	2. Questionnaire	3. Lab Observation	4. Questionnaire
1.	.90 (1)			
2.	.70 (2)	.90 (3)		
3.	-.30 (4)	0 (5)	.90 (6)	
4.	0 (7)	-.30 (8)	.70 (9)	.90 (10)

Here we have two hypothesized constructs important in early social development: fear of strangers and exploration in a novel environment. Each is measured using two methods, a laboratory session where observations of the child's behaviors are taken, and a maternal questionnaire. Let us assume a reasonably large sample size.

Reliability

The correlations on the main diagonal (cells 1, 3, 6, 10) are short-term stability coefficients. In this simplified sample, all short-term stability coefficients are .9, certainly high enough for one to feel confident about accuracy of measurement. Researchers using observational instruments, however, frequently have not thoroughly confronted issues of reliability (Rowley, 1976). Often only observer agreement indices are used as the sole measures of reliability when the stability of children's behaviors from one measurement to the next is the more fundamental issue.

There is some confusion concerning the distinction between observer agreement and stability (McGaw, Wardrop, and Bunda, 1972). Although they are both measures of reliability, they are not synonymous. Medley and Mitzel (1963) point out that intra- and inter-observer agreement are limiting factors on short-term stability. The extent to which observers disagree with themselves or each other limits the accuracy of measuring children's behaviors. Variability of the children under observation is the most important source of variance, and is the measure of stability. Short-term stability, therefore, deals with variability in the child and observer agreement deals with ob-

server consistency. This often-ignored distinction must be kept in mind. Observers may have high intra- and inter-observer agreement statistics, but may nevertheless collect data that does not show stability for at least two reasons: (1) differences among children may be too small, or (2) the behaviors sampled may be unstable.

Observer Agreement

There are two essential types of observer agreement that should be calculated. Intra-observer agreement estimates how consistent an observer is with himself or herself. It is usually determined by having each observer score the same videotapes or protocols on at least two separate occasions, and then statistically comparing the two sets of scores. Inter-observer agreement measures the extent to which at least two observers score a given situation similarly. It is estimated by statistically comparing ratings given by different observers of the identical situation.

Observer agreement indices carry a wealth of information. A thorough investigation and documentation of observer agreement is especially useful as a guide to other users of the instrument. Rough spots can be ironed out with each successive study and the instrument can be improved. Inter-observer agreement results are typically the only ones reported, however, neglecting intra-observer agreement, is also important in that it limits inter-observer agreement and results from its calculation can suggest valuable observer training strategies. If an observer doesn't consistently agree with him or herself, he or she can't agree with others. Low intra-observer agreement may indicate that one observer is having trouble interpreting particular

behaviors. Calculating intra-observer agreement statistics will help pinpoint observer difficulties and behaviors that are ambiguous to the observer.

Although not typically done, the use of an expert observer in inter-observer agreement studies can also help insure both the accuracy and the validity of the observation procedure. This is particularly important when one is using an observational system with a fairly high level of inference. For example, the Ainsworth Strange Situation is a commonly used research instrument designed to measure quality of attachment of children (ages 1-2) to their mothers. In the early stages of the Stanford Day Care Study, the strange situation raters had little difficulty achieving high levels of inter-observer agreement. But, when they compared their ratings to those of a rater trained by the developer of the scales, they found substantial disagreement. The raters agreed well about what they saw; however, they were attending to the wrong behaviors. This experience is probably not unusual. Researchers should therefore take the time to check their observers' scores with those of researchers using the same instrument elsewhere.

Stability

Once adequate observer agreement has been reached, we may confront the stability issues. Stability estimates indicate the consistency of children's behaviors. In correlational terms, a measure of stability compares behaviors from two time samples in the same, or similar contexts. In variance terms, stability compares the variance between children with the variance within children.

The concept of reliability has its origins in classical measurement theory and has historically been applied almost exclusively to paper-and-

pencil tests. Traditionally, several options have been used to estimate the reliability of a measurement procedure: test-retest, internal consistency, and parallel forms. The application of these concepts to observational measures depends on the structure of the observational instrument. For example, the test-retest method may be appropriate for obtaining stability data for observations in a naturalistic setting such as a day care center. A series of observations may be taken over a short period of time where the results of the time samples may be compared. A test-retest procedure, however, may not be appropriate for laboratory observational procedures. During the retest, the child may remember the first testing procedure, and this memory may unpredictably alter the child's behavior. However, if the retest is delayed a period of time to minimize memory effects, then maturation effects enter the picture. Like memory, maturation may create unpredictable variance in the behaviors that would adversely affect the stability estimate.

One solution to the short-term stability issue for lab procedures may be the "parallel form" concept. An advantage of this over test-retest is that unwanted variance due to memory is reduced. For example, to decrease memory effects, a parallel lab situation may differ from the initial procedure in the following ways: The session may be held in a different building and thus in a different room; different, but similar equipment and toys may be used; different, but parallel tasks could be used; a different person could be experimenter. If children are run through this parallel procedure shortly after the first procedure, then short-term stability of behaviors may be estimated.

Returning to the MTMM matrix, the correlations in the diagonal cell then should represent short-term stability estimates which, as we have seen, depend on observer agreement. We shall next discuss construct validity, but it should

be noted that the jump from reliability to validity is not great. Campbell and Fiske state that reliability and validity are regions on a continuum. Reliability is the agreement between two maximally similar methods designed to measure the same construct. Validity is the agreement between two maximally different methods designed to measure the same constructs.

MTMM-Construct Validity

Formal validity assessment has rarely been done in developmental research. Frequently rationalization from theory or face validity of procedures is relied upon to justify measurement instruments and the psychological labels used to describe behaviors. At best, face validity is simply not enough, and it may lead to faulty assumptions and conclusions. Construct validation, which explicitly relates empirical data to theory, is sorely needed and the MTMM framework is particularly useful for assessing it. MTMM analyses also provide information on the relation between constructs and may suggest where modifications in the theories underlying constructs are needed. (See Cronbach and Meehl, 1955, for an introduction to construct validity.)

Child development researchers use many different methods of assessment even when attempting to assess the same construct. For any given construct, the different methods probably result in somewhat different findings. Similarly, variations within a data collection technique would create variance in results. For example, just as a lab assessment and a maternal questionnaire measuring a child's exploration behaviors would not agree totally, neither would results from a maternal questionnaire and a questionnaire completed by a substitute caregiver such as a day care provider. In looking at a construct, it is imperative to compare results of different measurement

procedures as well as look at its relationship to other constructs. We are now ready to look at the off-diagonal entries of our example.

The off-diagonal correlations provide evidence for what Campbell and Fiske have called convergent and discriminant validity, both of which are necessary for construct validation. Validation must be convergent, in the sense that relatively independent measurement procedures rank order children on the hypothesized construct in a similar fashion. Discriminant validation requires that relatively lower correlations be found between different constructs. In our example, convergent validity data is found in cells 2 and 9, where correlations of .7 exist between methods within each construct. This indicates that the constructs to a large extent are strong enough to be measured by different methods. Discriminant validity data is in cells 4 and 8, where correlations of -.3 are found between constructs within each method. In other words, fear and exploration are related in the hypothesized direction: children who fear strangers are less likely to explore in a novel environment. Also, cells 5 and 7 contain data on discriminant validity, where neither traits nor methods are shared and the zero correlations would be predicted.

The theories underlying constructs should enter into the interpretations of the MTMM matrix (Golding, 1977; Tesser and Krauss, 1976) and therefore the observed discriminant validity must be consistent with theory. In the proceeding example, the -.30 correlation between the two constructs was acceptable because some social development theories suggest a mild negative relationship. Suppose another hypothesized construct were included in the matrix, for example some measures of intelligence. Developmental theory would not necessarily predict a relationship between, say, measures of intelligence and fear of strangers. A correlation of -.30 found in this case, therefore,

would not indicate discriminant validity.

Theory also plays a role in convergent validity, which deals with method variance. The degree of cross-method convergence expected is itself an assumption enbedded in the theory underlying a given construct. For example, a developmental theorist might predict a rather strong relationship between the results from a lab assessment and a maternal questionnaire; thus, the .7 correlation found in the above example fits nicely into the theory. The theorist, however, would probably predict different behavioral patterns depending on whether the mother is visible to the child. A lower correlation (say .3) between measures taken when the mother is present and when she is absent should not cause one to reject the construct fear of stranger. Limited convergence is expected in this case. In fact, a larger correlation between behaviors in the mother-present, mother-absent situation would indicate that the theory might need modification.

No single study can establish construct validity. It would indeed be expensive to collect all the data needed in a complete, formal MTMM treatment which included all the relevant methods and constructs. Also, any single study is bound to have empty cells in its MTMM matrix, i.e., not all traits will be measured by all methods. Finally, few studies in child development have large enough sizes that and themselves to a complete MTMM assessment.

As a result, construct validation must involve relating and combining the work of various investigators. Combining results of different studies enables knowledge to accumulate and allows for public validation of developmental constructs. Investigators interested in a construct could note how other researchers measured the construct and what the results were. One could also note the other constructs measured. One should be able to in-

formally, but systematically, fill the cells of a MTMM matrix with data collected from other research studies. This procedure of information accumulation would be neatly facilitated if researchers would include at least one established technique among their measures of the construct of interest. By comparing scores on an established procedure with scores on a newly developed one, the new technique would be anchored to other research. This process is analogous to validity studies of intelligence tests in which scores from a newly developed test are correlated with those from an established test like the Binet.

The process of building MTMM matrices from the results of a series of studies could also be facilitated if researchers reported more detailed accounts of their methods. The paucity of behavioral detail in published work has made the refinement of developmental theories difficult. The details must include precise descriptions of the experimental procedure and the particular variables assessed. Lack of precise communication between researchers leads to multiple labelling of identical behaviors and the same label meaning different things. When this happens, results are confusing and theory underlying constructs cannot expand. When the same labels are used for the same behaviors, when different labels are used for different bahaviors, and when the experimental context is clearly taken into account, researchers can build upon one another's efforts and validate their observational instruments and developmental constructs.

A few comments should be made about statistical analysis of these matrices. Traditionally, path analysis, factor analysis, and analysis of variance techniques have been applied to MTMM matrices. For most single studies with empty cells and small sample sizes, however, these techniques, and even

less demanding methods (e.g. Hubert and Baker, 1978), probably will not apply. For now, researchers should rely on simple observation of the correlations and interpret them in light of theory. On the other hand, a methodology called meta-analysis is being developed in part to combine correlations from different, but related studies (Glass, 1976). In doing so, this methodology may account for the quality, sample size, and similarity of measurement procedures of each study. This promising statistical orientation may be useful in combining results of different studies using observational procedures in a MTMM framework.

Conclusion

In conclusion, child development researchers need to attend to the psychometric qualitites of their observational instruments. They should assess not only observer agreement, but also stability and construct validity. The MTMM framework is useful in guiding the collection of psychometric data, particularly convergent and discriminant validity. These psychometric data should be interpreted in light of the theory underlying the constructs of interest. When researchers share these data with each other, knowledge may accumulate so as to lend support or nonsupport for the validity of their observational instruments and developmental constructs.

REFERENCES

1. Campbell, D.T., and Fiske, D.W. Convergent and divergent validation by the multitrait multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

2. Cronbach, L.J., and Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

3. Frick, T. and Semmel, M.I. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 1978, 48, 157-184.

4. Glass, G.V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

5. Golding, S.L. Method variance, inadequate constructs, or things that go bump in the night? Multivariate Behavioral Research, 1977, 12, 89-98.

6. Hubert, L.J. and Baker, F.B. Analyzing the multitrait-multimethod matrix. Multivariate Behavioral Research, 1978, 13, 163-179.

7. McGraw, B., Wardrop, J.L. and Bunda, M.A. Classroom observational schemes: where are the errors? American Educational Research Journal, 1972, 9, 13-27.

8. Medley, D.M., and Mitzel, H. Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago, Ill.: Rand-McNally, 1963, 247-328.

9. Rowley, G.L. The reliability of observational measures. American Educational Research Journal, 1976, 13, 51-59.

10. Tesser, A. and Krauss, H. On validating a relationship between constructs. Educational and Psychological Measurement, 1976, 36, 111-121.